Homework 2 Report - Income Prediction

學號：r06323011 系級：經濟碩一 姓名：葉政維

1. **(1%) 請比較你實作的 generative model、logistic regression 的準確率，何者較佳？**

Logistic 的表現較佳，使用幾乎一樣（generative model 去除 intercept）的 features，generative 在 private board 上的分數為 0.84166，而 logistic 則有 0.85530。generative model 假設 features 來自聯合常態分佈，但本筆資料多為類別變數，且除了教育之外，各類別間大多無序。在違反淚後假設的情形下，generative 的表現可能因此較差。

2. **(1%) 請說明你實作的 best model，其訓練方式和準確率為何？**

我僅使用 Logistic Regression 搭配 L2 regularizer。首先，部份變數有很多種的類別，有些類別的 size 非常小，而有些類別則相當類似。我決定先將部分的類別進行合併，合併依據是依照不同類別在 income 上的 unconditional mean，如果經過檢定後的 t-test 顯著，則進行合併。此外，工時此一變數的分佈有些奇妙，許多值集中在 5 的倍數附近，因此我將此一變數以 5 的變數附近為區間，進行間斷化；capital 此一變數多為 0，因此我另外增加 capital gain(loss)==0 的 dummy。

在 features 上，我圍繞著性別此一變數和其他變數去產生交乘項，對於連續變數，則嘗試增加二次、三次項，並以 5-fold CV 決定是否加入這些變數以及決定 regularizer。最終 private 上的準確率是 0.85530（其他較好的結果有到 0.85861，很可能是我過度以 public 來決定模型而 overfit public board）。

3. **(1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關 normalization 請參考：https://goo.gl/XBM3aE)**

首先，若不進行標準化，很多時候 Gradient Descend 會 overflow。為方便比較，features 將不包含 capital（值太大了，而包含的連續變數僅有 age、age^2），此外 learning rate 也有調整過以避免 overflow。

結果顯示：無標準化(learn=0.002)的 private accuracy 為 0.7655；標準化(learn=0.002)的 private accuracy 為 0.81599；標準化(learn=0.2)的 private accuracy 為 0.85431；即使在同一個 learning rate 下，標準化對於 Gradient Descend 進行極大化的過程仍相當大的幫助。依課堂上的解釋，乃是因為未標準化時，Gradient Descend 在某些方向的移動太慢了（本例中的 age^2）。

4. **(1%) 請實作 logistic regression 的正規化(regularization)，並討論其對於你的模型準確率的影響。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)**

以最終使用的 features 為基準（註：iteration 為 5000，而最終版本為 10000），分別比較 regularizer 為 0、0.5、3、5。private accuracy 分別為 0.8537、0.8548、0.85824、0.85775。加入適當的 regularizer 可以讓模型較為平滑，即不要對於某些變數過度反應，就結果來看確實有一定的幫助。

5. **(1%) 請討論你認為哪個 attribute 對結果影響最大？**

以沒有加任何交乘項的模型(regularizer 為 0)為基準，分別探討將哪個（組）變數移除會導致最多 accuracy 上的損失，未移除變數時的模型 private accuracy 為 0.85591。去除 age 為 0.8424；去除 workclass 為 0.85616；去除 education 為 0.84596；去除 marital status 為 0.85591；去除 occupation 為 0.84964；去除 relationship 為 0.85579；去除 race 為 0.85616；去除 sex 為 0.85517；去除 capital 為 0.83933；去除 country 為 0.85505；去除 work hour 為 0.8548。

結果顯示 education、occupation、capital(gain and loss)是重要的解釋變數，除去後 accuracy 差異不大的變數不一定不重要，很可能是已經被其他變數給解釋了，如 sex 可能有一部分被 relationship 所解釋。